

WHAT IS CLAIMED IS:

1 1. A method of extracting data from a file, the method comprising:
2 receiving a request to extract one or more data records from the file;
3 identifying the data records within the file, without using prior knowledge of the
4 structure of the file; and
5 extracting the data records.

1 2. The method of claim 1, wherein the data records contain at least one tag and text,
2 and wherein the method further comprises the step of separating each tag from the text by adding
3 a space between the tag and the text.

1 3. The method of claim 2, further comprising converting the text into a list that
2 contains one or more words, wherein each word is associated with its position in the list.

1 4. The method of claim 3, further comprising:
2 evaluating each word based on a meaning of the word and based on the position of the
3 word in the list; and
4 based on the evaluation, calculating a score for the word.

1 5. The method of claim 3, further comprising:
2 converting the text into a list that contains one or more phrases;
3 determining whether one of the words, at a specified position in the word list, matches one
4 of the phrases in the phrase list; and
5 based on that determination, calculating a score for the phrase.

1 6. The method of claim 3, further comprising:
2 determining whether one of the words, at a specified position in the word list, matches a
3 predefined expression; and
4 based on that determination, calculating a score for the predefined expression.

00728600-120100
001037-6892400

5 7. The method of claim 6, wherein the predefined expression is a dollar value.

1 8. The method of claim 4, further comprising:
2 identifying a region of interest within the word list by determining which region has a
3 highest grade; and
4 using the identified region of interest to create a set of sub-regions of the word list.

1 9. The method of claim 8, further comprising partitioning one of the sub- regions into
2 one or more record regions, wherein each record region contains text that corresponds to a single
3 record.

1 10. The method of claim 9, further comprising extracting one or more attribute values
2 from the record region.

1 11. The method of claim 10, wherein extracting the attribute values comprises
2 extracting one or more phrases.

1 12. The method of claim 10, wherein extracting the attribute value comprises
2 extracting a dollar amount

1 13. The method of claim 10, further comprising writing a set of records to a file.

1 14. The method of claim 1, further comprising writing the extracted data record to an
2 output file.

1 15. An apparatus for extracting data from a file, comprising:
2 a computer; and
3 a computer program, performed by the computer, for receiving a request to extract one
4 or more data records from the file, identifying the data records within the file, without using prior

5 knowledge of the structure of the file, and extracting the data records.

1 16. The apparatus of claim 15, wherein the data records contain at least one tag and
2 text, and wherein the apparatus further comprises a computer program, performed by the
3 computer, for separating each tag from the text by adding a space between the tag and the text.

1 17. The apparatus of claim 16, further comprising a computer program, performed by
2 the computer, for converting the text into a list that contains one or more words, wherein each
3 word is associated with its position in the list.

1 18. The apparatus of claim 17, further comprising:
2 a computer program, performed by the computer, for evaluating each word based on a
3 meaning of the word and based on the position of the word in the list, and based on the
4 evaluation, calculating a score for the word.

1 19. The apparatus of claim 17, further comprising:
2 a computer program, performed by the computer, for converting the text into a list that
3 contains one or more phrases, determining whether one of the words, at a specified position in
4 the word list, matches one of the phrases in the phrase list, and based on that determination,
5 calculating a score for the phrase.

1 20. The apparatus of claim 17, further comprising:
2 a computer program, performed by the computer, for determining whether one of the
3 words, at a specified position in the word list, matches a predefined expression, and based on that
4 determination, calculating a score for the predefined expression.

1 21. The apparatus of claim 20, wherein the predefined expression is a dollar value.

1 22. The apparatus of claim 18, further comprising:
2 a computer program, performed by the computer, for identifying a region of interest

3 within the word list by determining which region has a highest grade, and using the identified
4 region of interest to create a set of sub-regions of the word list.

1 23. The apparatus of claim 22, further comprising a computer program, performed by
2 the computer, for partitioning one of the sub- regions into one or more record regions, wherein
3 each record region contains text that corresponds to a single record.

1 24. The apparatus of claim 23, further comprising a computer program, performed by
2 the computer, for extracting one or more attribute values from the record region.

1 25. The apparatus of claim 24, wherein extracting the attribute values comprises
2 extracting one or more phrases.

1 26. The apparatus of claim 24, wherein extracting the attribute value comprises
2 extracting a dollar amount

1 27. The apparatus of claim 26, further comprising a computer program, performed by
2 the computer, for writing a set of records to a file.

1 28. The apparatus of claim 15, further comprising a computer program, performed by
2 the computer, for writing the extracted data record to an output file.

1 29. An article of manufacture comprising a computer program carrier readable by a
2 computer and embodying one or more instructions executable by the computer to perform the
3 method of extracting data from a file, the method comprising:

4 receiving a request to extract one or more data records from the file;
5 identifying the data records within the file, without using prior knowledge of the
6 structure of the file; and
7 extracting the data records.

1 30. The article of manufacture of claim 29, wherein the data records contain at least
2 one tag and text, and wherein the article of manufacture further comprises the step of separating
3 each tag from the text by adding a space between the tag and the text.

1 31. The article of manufacture of claim 30, further comprising converting the text into
2 a list that contains one or more words, wherein each word is associated with its position in the
3 list.

1 32. The article of manufacture of claim 31, further comprising:
2 evaluating each word based on a meaning of the word and based on the position of the
3 word in the list; and
4 based on the evaluation, calculating a score for the word.

1 33. The article of manufacture of claim 31, further comprising:
2 converting the text into a list that contains one or more phrases;
3 determining whether one of the words, at a specified position in the word list, matches one
4 of the phrases in the phrase list; and
5 based on that determination, calculating a score for the phrase.

1 34. The article of manufacture of claim 31, further comprising:
2 determining whether one of the words, at a specified position in the word list, matches a
3 predefined expression; and
4 based on that determination, calculating a score for the predefined expression.

1 35. The article of manufacture of claim 34, wherein the predefined expression is a
2 dollar value.

1 36. The article of manufacture of claim 32, further comprising:
2 identifying a region of interest within the word list by determining which region has a
3 highest grade; and

